



Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan

Frederick Jelinek¹

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 21218, USA

Received 24 January 1996

1. Homage à Bourlard

The article which has inspired these speculations [1] is a somewhat tame version of Bourlard's keynote address given at the 1995 Eurospeech meeting in Madrid [2]. The toning down is unfortunate but inevitable: no text could reflect Bourlard's personality and the zest of his presentation.

By encouraging the authors to be technically specific, the printed medium distorts the balance of the original message which was a spirited appeal to researchers and sponsors. Researchers were urged to exhibit courage and perseverance in the face of necessary difficulties and disappointments, and sponsors entreated to curb their understandable desire for a fast pay-off.

In his lecture Bourlard gave examples of his (and his co-workers') own courage and perseverance. The talk was imbued with irony and levity just like the quotation of Montaigne which graces the introduction to the paper.

It is often said approvingly of someone that while he may not have provided the solution, he asked the right questions. Most of the specific questions contained in the article [1] are open to serious challenge that others will no doubt take up. Nevertheless, the contribution of the authors is considerable: they call our attention to fundamental problems of research in *any* technology.

2. Speculation 1: HMMs

While we should always be on the lookout for a better formulation, it is quite probable that hidden Markov models (HMMs) are like the wheel, the building crane, or silicon: all perfect machines for their tasks. How many different implementations of memories have been suggested, each with great hopes? Josephson junctions, bubbles, etc. And yet, silicon seems an ever-improving moving target.

Even the great Shannon singled HMMs (he called them Markov sources) out for his attention in his paper that laid the foundations of Information Theory (see, for instance, Fig. 4, Section 5 in [3]).

In the keynote address at Snowbird [4] Steve Young advocated the state-of-the-art HMMs that use Gaussian mixtures. He pointed out their relative robustness, the convenience of a parametric approach, their easy applicability to adaptation.

Boullard, Hermansky, and Morgan, too, employ the HMM structure in their research, even though they use artificial neural networks (ANNs) to estimate the basic parameters.

The most important charge against HMMs is the conditional (on the state sequence) independence of outputs. There is no question that HMMs estimate absolute probabilities (densities!) $P(\mathbf{A} | \mathbf{W})$ very badly: just try to generate acoustic strings \mathbf{A} by HMMs! Yet the relative ratios

$$\frac{P(\mathbf{A} | \mathbf{W})}{P(\mathbf{A} | \mathbf{W}')}$$

¹ E-mail: jelinek@jhu.edu

between two alternative hypotheses \mathbf{W} and \mathbf{W}' may well provide a sufficiently accurate approximation for a choice between them.

The true weakness in our use of HMMs lies in the fact that we don't know how to induce their structure automatically from data, or what kinds of HMM building blocks to choose. The Baum–Welch algorithm allows us to estimate given model parameters, but not what parameters a model ought to be based on!

3. Speculation 2: The famous fudge factor

As many before them, Bourlard, Hermansky, and Morgan are concerned with the “mystery” of the employment of the fudge factor (known as the *balance parameter* in more polite circles) γ in the recognition criterion

$$\hat{\mathbf{W}}(A) = \arg \max_{\mathbf{W}} P_{\Theta}(A | \mathbf{W})^{\gamma} P_{\Phi}(\mathbf{W}) \quad (1)$$

on which the operation of practical large vocabulary recognition systems is invariably based. In (1) P_{θ} is a density that depends on the choice of a parameter set Θ whose value is determined from transcribed training data. P_{ϕ} is a probability distribution over word strings, the value of the parameter set Φ being selected on the basis of a training text. As the authors point out, the structure and the values of Θ and Φ are selected independently and that is surely a problem.

But let us get back to basics. With $\delta(\cdot, \cdot)$ denoting the Kronecker delta function, \mathbf{W} the spoken and $\hat{\mathbf{W}}(A)$ the recognized utterance (when the acoustics \mathbf{A} are observed), the commonly agreed on recognition criterion seeks to minimize the expected utterance recognition error

$$\begin{aligned} & E\{[1 - \delta(\mathbf{W}, \hat{\mathbf{W}}(A))] | A\} \\ &= \sum_{\mathbf{W}} [1 - \delta(\mathbf{W}, \hat{\mathbf{W}}(A))] P(\mathbf{W} | A) \\ &= 1 - \sum_{\mathbf{W}} \delta(\mathbf{W}, \hat{\mathbf{W}}(A)) P(\mathbf{W} | A) \\ &\geq 1 - \max_{\mathbf{W}} P(\mathbf{W} | A). \end{aligned} \quad (2)$$

The lower bound holds with equality if and only if

$$\hat{\mathbf{W}}(A) = \arg \max_{\mathbf{W}} P(A | \mathbf{W}) P(\mathbf{W}). \quad (3)$$

Note now that the difference between (1) and (3) does not lie only in that the former includes the parameter γ , but also in the presence of the parameters Θ and Φ ! The product $P(A | \mathbf{W}) P(\mathbf{W})$ in (3) refers to the *true* statistics of the utterance–recognition process, while $P_{\theta}(A | \mathbf{W})$ and $P_{\phi}(\mathbf{W})$ are just (separate!) estimates, and the ones we actually use are known to be very bad!

Furthermore, for practical recognizers the value of γ in (1) is chosen to minimize not the expected utterance error (2), but WER, the *per word* error rate.

What is the exact formula specifying the latter? Let $L(\mathbf{W}, \hat{\mathbf{W}})$ denote the Levenshtein distance [5] between the strings \mathbf{W} and $\hat{\mathbf{W}}$, that is, the minimum number of symbols that must be deleted from, inserted into, or substituted for the symbols of \mathbf{W} in order to *transform* this string into the string $\hat{\mathbf{W}}$ ² In fact, every large vocabulary recognizer (the state-of-the-art for SWITCHBOARD is 50% WER!) is *trying* to minimize

$$E\{L(\mathbf{W}, \hat{\mathbf{W}}(A)) | A\} = \sum_{\mathbf{W}} L(\mathbf{W}, \hat{\mathbf{W}}(A)) P(\mathbf{W} | A). \quad (4)$$

But a simple formula for $\hat{\mathbf{W}}(A)$ (analogous to (3)) that would minimize (4) does not exist, and it will surely take considerable effort before anyone comes up with even a decent approximation.³

4. Speculation 3: The way things were

I think that many who these days deplore the competitive atmosphere in which speech recognition

² It is easily seen that the $L(\cdot, \cdot)$ function is symmetrical, that is, that $L(\mathbf{W}, \hat{\mathbf{W}}) = L(\hat{\mathbf{W}}, \mathbf{W})$.

³ That (3) does not minimize (4) can be seen from the following simple counterexample. Let $\mathbf{W} \in \{00, 01, 11\}$, with $P(\mathbf{W} | A) \in \{\frac{1}{3} + \epsilon, \frac{1}{3} - \frac{1}{2}\epsilon, \frac{1}{3} - \frac{1}{2}\epsilon\}$, in that order. Then choosing $\hat{\mathbf{W}} = 00$ gives an expected Levenshtein distance of $1 - \frac{3}{2}\epsilon$, while choosing the less probable $\hat{\mathbf{W}} = 01$ will give only $\frac{2}{3} + \frac{1}{3}\epsilon$.

research is being carried out are quite unaware of how things stood at the dawn of the modern ASR era, in the nineteen seventies. It is no doubt true that the currently frequent run-offs put a premium on minor system modifications and discourage bold innovation. But we cannot do without run-offs completely: we must seek a happy medium.

In the seventies, speech recognition was under the shadow of an article by J.R. Pierce [6] that asserted that the field was in the hands of charlatans who promised a lot but faked their results, and that the ASR problem is not amenable to a satisfactory solution in any case. Pierce was a very famous scientist from Bell Laboratories with great accomplishments in communications research, a person respected by the likes of Shannon, a grand old man finishing up his illustrious career at Cal Tech. And he was not altogether wrong!

In those days, researchers in our field would not distinguish between training and test data. Mathematics was absent. Difficulty of tasks was measured by the *branching factor*, that is, by the size of the biggest branch in the usual artificial grammar that produced the sentences to be recognized⁴. Every recognizer component was hand-crafted. Recognition decisions were based on *scores* attached by experts to “knowledge sources”.

In the seventies, various phonetic, acoustic, and linguistic experts were assigned to most teams. They joined them reluctantly, having more respectable things to do. Management as well as sponsors (this was during the first ARPA project that ended in 1976) listened to their opinions. They thought that “if only engineers were willing to learn the basics ...”.

When I inquired of one of our assigned linguists what we should do when we move beyond the artificial syntax of the Raleigh grammar [7] we were using, he told me not to worry, that he’ll just “write a little grammar of English ...”⁵.

The absence of citations documenting the asser-

tions I make here is not accidental. Why embarrass anyone?

5. Divertimento Ronald Cole

During one of the discussion periods at Snowbird, Ron Cole asked for the floor and urged everyone to concentrate on those applications of speech recognition which enhance “the quality of life” [8]. I think he mainly had in mind the various telephone or personal computer applications that are slowly finding their way into the market.

But after granting that speech recognition will be (and already is) of great help to the handicapped, I would like to sound a word of caution before we jump to any conclusions as to which products are likely to enhance (everyday) life. Because with the exception of credit and debit cards and advances in medicine which touch practically everyone, electronic technology directly improves only productivity, and often lowers costs by simply shifting them from the provider to the user. Indeed, speech recognition is sold on the basis of its potential to displace operators, but that will not enhance *their* quality of life, while the net effect on the customer may well be negative.

I think that very large vocabulary recognizers may have a much better chance to at least do no harm. For instance dictation machines, or foreign language teachers. And what about defence sponsored work on the recognition of conversational speech? Always supposing that government *stays in the hands of decent people* [9], won’t interdiction of terrorists and drug smugglers tend to *preserve* the quality of life, if not enhance it?⁶

⁶ Note (for non-American readers of *Speech Communication*): In the United States large-vocabulary conversational speech recognition is supported by the National Security Agency. In a recent series of articles [10] it has been pointed out that:

- NSA agents overheard Cali operatives plotting to smuggle 12,250 kilos of cocaine to Florida inside concrete posts.
- The agency reported the missile site that would shoot down Capt. O’Grady, though the information was not acted upon in time.
- NSA computers searched the airwaves for the terrorist Carlos for two decades before picking up information that led to his arrest in the Sudan.

⁴ That is why we could not use the entropy measure and had to invent the concept of *perplexity* so as to find some common ground with the branching factor crowd.

⁵ Could this be the reason for the folkloric: “Any time I fire a linguist, the recognition rate improves”?

But attempts at benefit analysis of research matter very little. With very rare exceptions, research is neither market driven nor welfare driven, but ego driven. The motivation of the practitioners of science, as of art, is self-realization.

6. Speculation 4: Can we do without statistics

In Section 3 of their article, Bourlard, Hermansky, and Morgan say: “One can still imagine systems that are not statistical in nature, and given the appropriate theory, such approaches should be tried.”

Once it is granted that devices such as ANNs are statistical, then what would an “appropriate” non-statistical theory look like? Statistics are an indispensable approach to the modeling of variability and of ignorance about its causes: we will always be ignorant. Statistics are also an evaluation tool reflecting our faith that the future will be like the past. But chaos (in its colloquial, non-technical meaning) seems the only alternative to such faith.

It isn't that statistics is a religion. It is that nothing better is known ... Statistics wins because its techniques are well developed. “Knowledge” loses because those who think they have it do not know how to incorporate what they have. So far, it has turned out more profitable to estimate reliably very simple parameters than to introduce a complex model whose parameter values cannot be verified.

Airplanes do not flap wings but have wings nevertheless, says the article [1] in Section 4.1.2. Of course, we should try to incorporate the knowledge that we have of hearing, speech production, etc., into our systems, but first we must figure out how to parametrize it, and how to estimate the parameter values from speech data. There is no other way.

Perhaps the major barrier to progress in our field is the fact that language phenomena are Zipf-like so that their tail is very large and cannot be captured by any crude statistics. These phenomena must first be re-mapped into a space in which their Zipf nature is alleviated if it won't disappear altogether. In any case, we must reconcile ourselves with the sad fact that we will always be dealing with the sparse data problem.

Which means smoothing of estimates. And we have no satisfactory theory of smoothing *even* for

the unrealistic case that test data will be very much like training data! This has been driven home to me with particular force when we recently attempted to construct a decision tree language model that predicts letters. The usual interpolated *n-gram* model is a special case of a decision tree model, and yet, as much as we tried, we could not invent a principled smoothing algorithm that would allow a general decision tree model (which has much more freedom!) to beat the *n-gram* model, even if the former was so restricted that the latter was embedded in it.

Finally, a statistical approach involves the use of optimization criteria. We know how to apply MLE and MMI, but both these criteria are indirect and involve a lot of faith. After all, what we *really* seek is the lowering of the error rate! It is by now well known that the performance of HMM part-of-speech taggers grows worse with each iteration of the Baum-Welch algorithm when the *initial* parameter values are those obtained from counts of a hand-tagged training corpus [11]. And, on the other hand, we have noted the remarkable success of the simple transformational tagger [12] whose transformations are selected according to a *direct* error rate criterion.

7. Speculation 5: An immodest proposal: speech recognition as a code-breaking activity

Practical automatic speech recognition is of necessity a (near) real time procedure performed by a system whose structure is fixed and whose parameters once trained may be adapted on the basis of the speech that the system observed during recognition.

However, it would be at least of *academic* interest to view the recognition task as a code-breaking activity to whose accomplishment can be devoted an essentially *infinite* amount of effort. From this point of view we could ask:

How well can we do if, in order to transcribe a single sentence, we are allowed to completely retrain all models, even on the basis of data acquired specifically for the processing of the particular sentence at hand?

Under code-breaking rules almost everything would be fair, including, for instance, (a) the retraining of a language model on the basis of newly

acquired text, (b) the retraining of the acoustic model on the basis of newly recorded and transcribed speech, or, (c) a complete change of the recognizer paradigm.

The only thing forbidden would be any human intervention, such as listening to the speech to be recognized, viewing its spectrogram, etc.

References

- [1] H. Bourlard, H. Hermansky and N. Morgan (1996), “Towards increasing speech recognition error rates”, *Speech Communication*, in this issue.
- [2] H. Bourlard (1995), “Towards increasing speech recognition error rates”, *Proc. Eurospeech'95, Madrid*, pp. 883–894.
- [3] C.E. Shannon (1948), “A mathematical theory of communication”, *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656.
- [4] S. Young (1995), “Large vocabulary continuous speech recognition”, *Proc. 1995 IEEE Automatic Speech Recognition Workshop, Snowbird, UT*, pp. 3–29.
- [5] V.I. Levenshtein (1966), “Binary codes capable of correcting deletions, insertions and reversals”, *Soviet Physics Dokl.*, vol. 10, no. 10, pp. 707–710.
- [6] J.R. Pierce (1969), “Whither speech recognition?” *J. Acoust. Soc. Amer.*, vol. 46, pp. 1049–1051.
- [7] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer and H.F. Silverman (1976), “Preliminary results on the performance of a system for the automatic recognition of continuous speech”, *1976 Internat. Conf. Acoust. Speech Signal Process., Philadelphia, PA*.
- [8] R. Cole (1995), Remarks made at the *1995 IEEE Automatic Speech Recognition Workshop, Snowbird, UT*.
- [9] K. Capek (1935), *President Masaryk tells his story* (G.P. Putnam, New York).
- [10] S. Shane and T. Bowman (1995), “America’s fortress of spies”, *The Baltimore Sun*, Dec. 3–15, 1995.
- [11] B. Merialdo (1984), “Tagging English text with a probabilistic model”, *Computational Linguistics*, vol. 20, no. 2, pp. 155–172.
- [12] E. Brill (1996), “Learning to parse with transformations”, In: H. Bunt and M. Tomita, eds., *Recent Advances in Parsing Technology* (Kluwer Academic Publishers, Dordrecht).